Mol Gen Genomics (2006) 276: 1–12
DOI 10.1007/s00438-006-0122-z

ORIGINAL PAPER

**Chun Wan J. Lai · Qingyi Yu · Shaobin Hou
Rachel L. Skelton · Meghan R. Jones
Kanako L. T. Lewis · Jan Murray · Moriah Eustice
Peizhu Guan · Ricelle Agbayani · Paul H. Moore
Ray Ming · Gernot G. Presting**

Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome

Received: 18 February 2006 / Accepted: 22 March 2006 / Published online: 16 May 2006

Abstract Papaya (*Carica papaya* L.) is a major tree fruit crop of tropical and subtropical regions with an estimated genome size of 372 Mbp. We present the analysis of 4.7% of the papaya genome based on BAC end sequences (BESs) representing 17 million high-quality bases. Microsatellites discovered in 5,452 BESs and flanking primer sequences are available to papaya breeding programs at http://www.genomics.hawaii.edu/papaya/BES. Sixteen percent of BESs contain plant repeat elements, the vast majority (83.3%) of which are class I retrotransposons. Several novel papaya-specific repeats were identified. Approximately 19.1% of the BESs have homology to *Arabidopsis* cDNA. Increasing numbers of completely sequenced plant genomes and BES projects enable novel approaches to comparative plant genomics. Paired BESs of *Carica*, *Arabidopsis*, *Populus*, *Brassica* and *Lycopersicon* were mapped onto the completed genomes of *Arabidopsis* and *Populus*. In general the level of microsynteny was highest between closely related organisms. However, papaya revealed a higher degree of apparent synteny with the more distantly related poplar than with the more closely related *Arabidopsis*. This, as well as significant colinearity observed between peach and poplar genome sequences, support recent observations of frequent genome rearrangements in the *Arabidopsis* lineage and suggest that the poplar genome sequence may be more useful for elucidating the papaya and other rosid genomes. These insights will play a critical role in selecting species and sequencing strategies that will optimally represent crop genomes in sequence databases.

Keywords Bacterial artificial chromosome · *Carica papaya* · Comparative genomics · Microsatellite · Genome mapping

Abbreviations BAC: Bacterial artificial chromosome · BES: BAC end sequence · kb: Kilobase · Mbp: Megabase pairs · MYA: Million years ago · nt: Nucleotide · SSR: Simple sequence repeat

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00438-006-0122-z and is accessible for authorized users.

Communicated by R. Hagemann

Chun Wan J. Lai and Qingyi Yu have contributed equally to this work.

C. W. J. Lai · M. Eustice · P. Guan · R. Agbayani · G. G. Presting (✉)
Department of Molecular Biosciences and Bioengineering,
University of Hawai'i, 1955 East-West Road,
Agricultural Sciences Building Room 218,
Honolulu, HI 96822, USA
E-mail: gernot@hawaii.edu
Tel.: +1-808-9568861
Fax: +1-808-9563542
URL: www.genomics.hawaii.edu/papaya/BES/

Q. Yu · R. L. Skelton · M. R. Jones · J. Murray · M. Eustice
P. Guan · R. Agbayani · R. Ming
Hawaii Agriculture Research Center, Aiea, HI 96701, USA

S. Hou · K. L. T. Lewis
Center for Genomics, Proteomics and Bioinformatics
Research Initiative, University of Hawai'i,
Honolulu, HI 96822, USA

P. H. Moore
USDA-ARS, Pacific Basin Agricultural Research Center,
Hilo, HI 96720, USA

R. Ming
Department of Plant Biology, University of Illinois at
Urbana-Champaign, Urbana, IL 61801, USA

Introduction

Papaya (*Carica papaya* L.) is a major tree fruit crop of tropical and subtropical regions, where it is grown primarily for the fresh fruit market. At only 372 Mbp (Arumuganathan and Earle 1991), the genome of papaya is at least 10% smaller than that of rice, the only completely sequenced crop plant to date. The small genome size of papaya and the fact that it can produce fruit in as

few as 9 months make it a potential model organism for fruit-producing tree crops (Ming et al. 2001). An added incentive to analyze this particular plant genome is provided by the recent identification of a primitive sex chromosome in papaya (Liu et al. 2004), which has commercial implications, as hermaphrodites have preferred agronomic characteristics.

BAC libraries, constructed with vectors that can stably hold inserts of up to 300 kb, have become important tools in crop genomics and have been used to construct physical maps (Chen et al. 2002), map genes of agricultural importance (Lange and Presting 2004), perform comparative genomics between crop species (O'Neill and Bancroft 2000; Ilic et al. 2003) and analyze genome structure by fluorescence in situ hybridization (Cheng et al. 2001). BAC libraries of many important crop genomes including rice, maize, tomato, sorghum, wheat and soybean have been constructed. Recently, Ming et al. (2001) reported construction of a papaya BAC library, containing 39,168 clones and estimated to represent 13.7 genome equivalents, from the Hawaiian papaya cultivar "Sun Up", using the restriction enzyme *HindIII* and the vector pBeloBAC11 (Shizuya et al. 1992).

BAC end sequence data are an important component of scaffolds used for the sequencing of large eukaryotic genomes. Physical maps constructed from BACs, together with associated end sequence information, can be used to sequence genomes by 'walking' from one clone to the next (Rice Chromosome 10 Sequencing Consortium 2003) or to anchor whole genome shotgun sequence data (Goff et al 2002). In addition, BAC end sequences provide a first glimpse of the sequence composition of an unsequenced genome (Mao et al. 2000; Zhao et al. 2001; Hong et al. 2004) and yield molecular markers useful for genetic mapping and breeding (Tomkins et al. 2004).

Here, we report the analysis of 50,661 BAC end sequences, which provide a first glimpse into the sequence composition of the papaya genome. Our analysis focuses on microsatellite contents, repeat element composition, protein-coding regions and comparative mapping of BAC-end sequence pairs to other sequenced plant genomes. The annotated BAC-end sequences will serve as useful resources for physical mapping, positional cloning, genetic marker development and genome sequencing of papaya.

Methods

Papaya BAC end sequencing

BAC DNA was isolated using R.E.A.L. Prep 96 (QIAGEN, CA, USA) according to manufacturer's guidelines, cycle sequenced with ABI BigDye Terminator v3.1 (ABI, CA, USA), and analyzed on ABI 3700 and ABI 3730 xl DNA Analyzers (ABI). Base calling of chromatograms and trimming of BAC end sequences were performed with PHRED software (Ewing et al. 1998; Ewing and Green 1998) using the default trimming cutoff value. PHRED output was converted into FASTA formatted sequence files using PHD2FASTA, and pBeloBAC11 cloning vector sequence was masked with CROSS_MATCH (http://www.genome.washington.edu). Terminal CROSS_MATCHed vector sequences were trimmed, and BESs shorter than 50 bp were eliminated. These non-redundant high-quality papaya BAC-end sequences were used in subsequent analyses.

Simple sequence repeats

For computational microsatellite detection, bases with PHRED quality value of less than 20 were converted to 'N's. Trimmed, high-quality papaya BAC-end sequences were scanned with all combinations of mono-, di-, tri-, and tetra-nucleotides to identify simple sequence repeats (SSRs). All SSRs spanning 12 or more nucleotides were recorded. Offset patterns (e.g. TATA and ATAT) were recorded only once based on repeat length and identity of the first base. PRIMER3 software (Rozen and Skaletsky 2000; code available at http://fokker.wi.mit.edu/primer3/) was used to design flanking PCR primer pairs with an optimal length of 20 bases, a GC percentage of 40–60% and a product size of 100–500 bp. If multiple SSRs were identified in a single BES, the SSR most useful for mapping was identified based on, in order of priority, (1) our ability to design flanking primers, (2) longest SSR and (3) minimum amplicon size.

Repeat analysis

After masking simple sequence repeats, BAC end sequences were compared with the The Institute for Genomic Research (TIGR) plant repeat databases (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/) containing 30,481 non-redundant annotated plant repeat sequences, using TBLASTX (Altschul et al. 1990) at an *E* value cutoff of 10^{-4}. A detailed listing of specific plant repeat databases used is available at http://www.genomics.hawaii.edu/papaya/BES/. BAC-end sequences were annotated using the best match in the repeat database and classified based on the TIGR *Codes for Plant Repetitive Sequences* table (http://www.tigr.org/tdb/e2k1/plant.repeats/repeat.code.shtml). To identify BACs derived from the papaya chloroplast, all high-quality BESs were screened with the *Arabidopsis thaliana* plastid genome (NC_000932.1) using BLASTN with an *E* value cutoff of 10^{-3} and requiring an alignment length of at least 100 nucleotides and an identity of $\geq 90\%$.

Annotation

High-quality BAC-end sequences with no homology to the repeat database and with SSR regions masked were compared to the August 26, 2004 *A. thaliana* cDNA database (TIGR, ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES) to identify BAC ends with homology to coding regions. BESs with matches in the *Arabidopsis*

cDNA database (TBLASTX with a cutoff value of 10^{-6}) were annotated with the original *A. thaliana* cDNA database annotation. These BAC-end sequences were further annotated by BLASTN comparison to the RefSeq plant genomic database (downloaded from ftp://ftp.ncbi.nih.gov/refseq/release/plant/) with a cutoff value of 10^{-6}. BESs without homology to sequences in any of the three databases (plant repeat, *Arabidopsis* cDNA or RefSeq) were annotated using the non-redundant protein database (*nr*) (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/) and BLASTX with an *E* value cutoff of 10^{-6} and an alignment length of at least 34 amino acids.

Papaya-specific repeat sequences

The most abundant papaya-specific repeats were identified by comparing BESs lacking detectable homology to the plant repeat database, *Arabidopsis* cDNAs, RefSeq or *nr* against each other using BLASTN with an *E* value cutoff of 10^{-10}. Up to 1,000 unique subject IDs were collected for each query and the BLAST results were imported into a database. The query with the largest number of matches (to other BESs) of at least 100 nt alignment length, and all its matches, were removed from the database. This procedure was repeated until the top 10 most frequent queries had been identified. These 10 queries were compared to the GenBank non-redundant nucleotide database with BLASTN via the internet-based server (http://www.ncbi.nlm.nih.gov/blast/) to identify homologous sequences. Queries without homology to any GenBank sequence and their matches were clustered using ClustalX (Thompson et al. 1997) in an attempt to identify consensus sequences.

Comparative genome mapping

Databases

Whole genome sequences of *A. thaliana* (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/), rice (*Oryza sativa*, ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_3.0/) and poplar (*Populus trichocarpa*, http://genome.jgi-psf.org/Poptr1/Poptr1.download.ftp.html) were downloaded from TIGR and the Joint Genome Institute and fragmented into 300 kb segments with 1,020 nucleotide overlap.

Queries

Low-copy BAC end forward and reverse pairs were selected from various BES datasets based on lack of homology to the repeat database, and SSRs contained in these BESs were masked. These paired "low-copy" BESs were mapped to each of the fragmented plant genomes using TBLASTX with an *E* value cutoff of 10^{-6}. If the highest scoring alignment of both the forward and reverse read were in the correct orientation and separated by at least 10 kb and not more than 300 kb in the target genome, the BAC was considered to be potentially colinear with the target genome.

All papaya BESs were included in this analysis. In addition, 12,000 *Arabidopsis* BESs were downloaded from the TIGR website (ftp://ftp.tigr.org/pub/data/a_thaliana/bac_end_sequences/atends), and 38,245 *Brassica rapa* and 50,000 *Lycopersicon esculentum* (tomato) BESs were downloaded from NCBI (http://www.ncbi.nlm.nih.gov, Brassica rapa [ORGANISM] BAC "end sequence", lycopersicon [ORGANISM] BAC pBeloBAC11). The poplar genome was fragmented into artificial BACs of 120 kb in length and separated by 10 kb gaps. Five-hundred nucleotides were collected from each end of these artificial BACs to generate 7,135 artificial poplar BAC end sequence pairs. Elimination of BESs with plant repeat homologies generated 9,038 papaya, 3,648 *Arabidopsis*, 8,525 *B. rapa*, 10,394 tomato and 5,685 poplar BES pairs.

A subset of 185 *Arabidopsis* BES pairs was used to validate the computational pipeline. SSRs were masked and high-copy BESs with homology to the repeat database were eliminated to yield 102 low-copy paired BAC end sequences, which were mapped to the *Arabidopsis* genome sequence using the same computational pipeline.

Results

BAC-end sequencing

A total of 50,661 BAC end sequence chromatograms were generated from 26,017 BAC clones from the library described by Ming et al. (2001). These BESs were base-called and trimmed using PHRED with default settings, yielding 39,590 high-quality BAC end sequences. CROSS_MATCH was used to mask vector sequences, and the masked terminal vector sequences were removed. A total of 1,548 sequences that consisted entirely of vector sequence, and 2,570 BESs shorter than 50 bases were eliminated to generate 35,472 high-quality sequences from 20,842 BAC clones. The total number of high-quality bases is 17,483,563 or 4.7% of the papaya genome. The trimmed sequences ranged from 50 to 899 nucleotides with an average of 493 bases. Eighty-seven percent of nucleotides had PHRED quality values ≥ 20. The G + C content of the BESs was 35%. All BESs of length ≥ 50 nucleotides were deposited to the GenBank GSS database (accession numbers DX458351–DX502755).

Plant repeat elements

Comparison of 35,472 high-quality BAC end sequences, in which all simple sequence repeats had been masked, to the plant repeat database revealed 5,733 (16.2%) end sequences with homology to plant repeat elements (Table 1). Class I retrotransposons represent the most abundant repeats with a total of 4,773 (83.3%). BAC-end

Table 1 Summary of plant repeat element content of 35,472 high-quality BESs

Class	Element	No.	Total
I (Retrotransposons)	*Ty1-copia*	1,425	4,773
	Ty3-gypsy	1,389	
	LINE	420	
	Unclassified retrotransposons	1,539	
II (Transposons)	CACTA, En/Spm	387	426
	Mutator (MULE)	9	
	Mariner (MLE)	1	
	Unclassified transposons	29	
III (Miniature inverted-repeat transposable elements)	Gaijin/Gaigin	1	9
	p-SINE1	2	
	MITE-adh, type D	3	
	Micron	2	
	Pangrangja	1	
Centromere-related sequences	Centromere-specific retrotransposons	220	242
	Unclassified centromere sequences	22	
Telomere-related sequences	Telomere associated	7	19
	Telomere	12	
Ribosomal RNA genes	45S rDNA	104	167
	5S rDNA	63	
Unclassified	Unclassified	97	97
	Total		5,733

BESs were compared with the TIGR plant repeat database and categorized based on the TIGR *Codes for Plant Repetitive Sequences*. The total number of BESs with homology to each class and repeat element is listed

sequences homologous to retrotransposons were further classified as long terminal repeat-containing *Ty1-copia* (24.9%) and *Ty3-gypsy* (24.2%), or LINE (7.3%); 1,539 (26.8%) of the retrotransposons could not be clearly assigned to a specific group (Table 1). The next most abundant repeats were 426 (7.4%) class II transposons. Only 9 MITEs were found. Unclassified plant repeat elements accounted for 1.7% (97) of all repeat elements. Other known plant repeat elements that could be identified in the papaya BES data include rDNA (2.9%) as well as centromere (242 = 4.2%) and telomere (19 = 0.3%) related repeats. A search of the high-quality BES set (obtained from 20,842 different BACs) with the *Arabidopsis* chloroplast genome identified 290 BESs from 230 different BACs (1.1%) as likely chloroplast-derived.

Coding regions

Only the 29,703 BESs without homology to the plant repeat database were included in the following analyses. Of these, 6,769 (representing 19.1% of the total high-quality BESs) were homologous to at least one *A. thaliana* cDNA. In 2,659 cases (39.3% of BESs with an *Arabidopsis* match) the best match was annotated as "putative", "unknown", "hypothetical", or "expressed protein". Homologies to RefSeq genomic sequences were found in 796 BESs (2.2%) with no homology to any *A. thaliana* cDNA. Of those BESs for which no homolog was identified in the plant repeat, *Arabidopsis* cDNA or RefSeq genomic databases, only 520 BESs (1.5%) had detectable homologs in the *nr* protein database.

Papaya-specific repeat sequences

BAC end sequences that had no homology in the TIGR plant repeat, *Arabidopsis* cDNA, RefSeq genomic or non-redundant protein database, were grouped based on homology. Examination of the 10 largest clusters revealed 6 that had clear homology to the *Carica papaya* male-specific region of the Y chromosome (MSY) sequences cpsm54 (gi|37992834), cpsm49 (gi|37992830) and cpbe55 (gi|37992870), which match 767, 161 and 120 BESs, respectively (Table 2). In addition, four previously unknown papaya-specific repeats that match 140, 116, 106 and 103 BESs were identified among these uncharacterized sequences.

Simple sequence repeats

A total of 7,456 SSRs of at least 12 nucleotides in length were identified in 5,452 (15.4%) BESs (Table 3). Of these SSRs, 1,174 (15.7%) span $\geq$ 20 nucleotides and thus represent hypervariable markers. The SSRs consist to 22.9, 37.8, 17.7, and 21.5% of mono-, di-, tri-, and tetranucleotide tandem repeats, respectively. Thus dinucleotide repeats are the most abundant class of microsatellites, followed by homopolymers. Poly(T) and poly(A) are the most abundant homopolymers, representing 882 (51.6%) and 736 (43.1%) of all homopolymers, respectively. While poly(AT) (1,137 or 40.3%) and poly(TA) (844 or 29.9%) are the most abundant dinucleotide repeats, poly(AG) and poly(GA) make up some of the longest microsatellites (Fig. 1). Poly(CG)/(GC) were rarely found.

Table 2 Annotation of the ten most abundant papaya-specific repeats

Repeat	BES	No.	Homology
1	30B-C12.r	305	gi|37992834 *C. papaya* isolate cpsm54 Y male-specific seq.
2	35C-A09.r	178	gi|37992834 *C. papaya* isolate cpsm54 Y male-specific seq.
3	47A-C11.r	177	gi|37992834 *C. papaya* isolate cpsm54 Y male-specific seq.
4	4B-F10.r	161	gi|37992830 *C. papaya* isolate cpsm49 Y male-specific seq.
5	53A-E11.f	140	Novel
6	44C-H12.r	120	gi|37992870 *C. papaya* isolate cpbe55 Y male-specific seq.
7	94B-G03.r	116	Novel
8	46C-C11.r	107	gi|37992834 *C. papaya* isolate cpsm54 Y male specific seq.
9	28D-B03.r	106	Novel
10	25C-H06.f	103	Novel

Repeats are numbered consecutively from most to least abundant *BES* BAC end sequence used as query, *No.* number of BESs with homology to query BES, *Homology* sequence accession number and description of closest GenBank homolog

Table 3 Distribution of SSRs in BAC ends with cDNA, repeat or no known homology (other)

	Total # SSR	cDNA (19.1%)	Repeat (16.2%)	Other (64.7%)
Mono				
Total	1,708	252 (14.8%)	90 (5.3%)	1,366 (80.0%)
A/T	1,618	246 (15.2%)	87 (5.4%)	1,285 (79.4%)
C/G	90	6 (6.7%)	3 (3.3%)	81 (90.0%)
Di				
Total	2,819	431 (15.3%)	105 (3.7%)	2,283 (81.0%)
AT/TA	1,981	289 (14.6%)	77 (3.9%)	1,615 (81.5%)
AG/GA	264	51 (19.3%)	3 (1.1%)	210 (79.5%)
CT/TC	286	41 (14.3%)	10 (3.5%)	235 (82.2%)
Other	288	50 (17.4%)	15 (5.2%)	223 (77.4%)
Tri				
Total	1,323	289 (21.8%)	80 (6.0%)	954 (72.1%)
AAT/ATA/TAA	269	27 (10.0%)	10 (3.7%)	232 (86.2%)
TTA/TAT/ATT	271	24 (8.9%)	20 (7.4%)	227 (83.8%)
Other	783	238 (30.4%)	50 (6.4%)	495 (63.2%)
Tetra				
Total	1,606	218 (13.6%)	97 (6.0%)	1,291 (80.4%)
AATT/ATTA/TTAA/TAAT	376	37 (9.8%)	12 (3.2%)	327 (87.0%)
Other	1,230	181 (14.7%)	85 (6.9%)	964 (78.4%)

BAC ends with cDNA or plant repeat homology represent 19.1 and 16.2% of the total, respectively. The percentage of each SSR type and selected patterns within these classes of BESs are shown (for more details please refer to http://www.genomics.hawaii.edu/papaya/BES/). Note: all SSRs are listed, even where multiple SSRs exist in a single BES

Poly(AAT) and poly(ATT) are the most abundant trinucleotide repeats, representing 12.2% (161) and 9.0% (119) of that class. The longest trinucleotide repeat is a poly(TTA) of 60 bases. Poly(AATT) and poly(AAAT) are the most abundant tetranucleotide repeats, representing 9.7% (156) and 8.1% (130) of that class (data not shown). AT-rich SSRs were consistently found to be more abundant than GC-rich SSRs.

Of the 5,452 BESs containing one or more SSRs, 16.9% (922) and 5.8% (318) had homology to *Arabidopsis* cDNAs or known plant repeats, respectively. Since 19.1 and 16.2% of all BESs have homology to cDNAs and plant repeats, respectively, it appears as though SSRs are slightly underrepresented in coding regions, very much underrepresented in elements of the plant repeat database and overrepresented in the remaining BESs (64.7%). Some trinucleotide repeats appear to be enriched in BESs with homology to cDNAs and may thus constitute better markers for gene-rich regions (Table 3).

Comparative genome mapping

After elimination of papaya BESs with homology to the plant repeat database, 9,038 BAC-end forward and

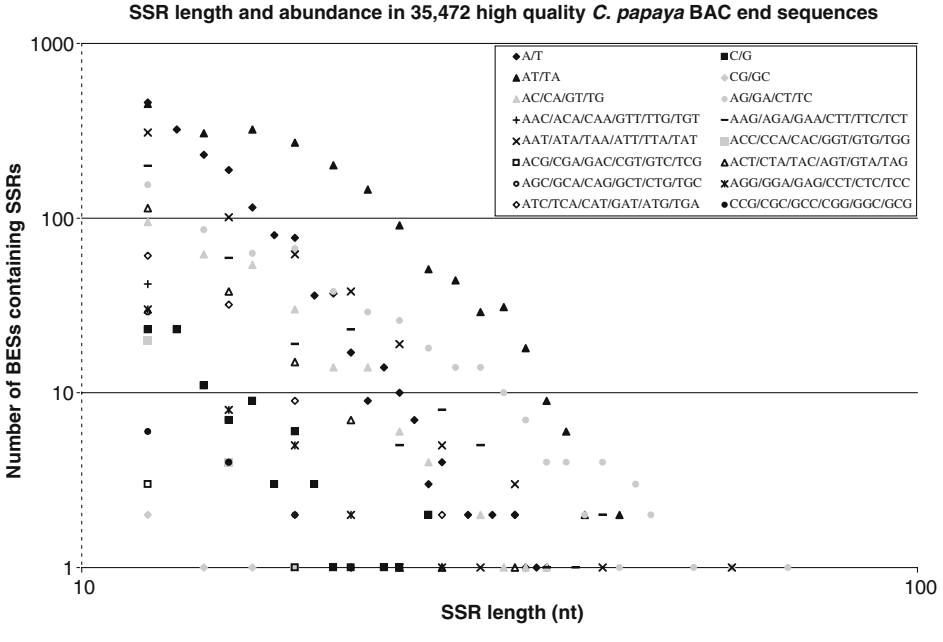

Fig. 1 Characterization of all microsatellites detected in 35,472 high-quality BESs. Numbers of BESs containing an SSR are plotted against the length for each microsatellite

reverse pairs remained. These paired BAC ends were subsequently mapped to the *A. thaliana*, *Populus trichocarpa* and *Oryza sativa* genomes. The paired BAC ends had to map between 10 and 300 kb in the heterologous genome and be oriented properly with respect to each other in order to be considered as potentially colinear with the target genome. These conditions were met by 53 (0.6%), 167 (1.8%) and 11 (0.1%) of the papaya BAC end sequence pairs in the *Arabidopsis*, poplar and rice genomes, respectively (Table 4).

The *Populus* genome not only yielded the most mapped papaya BAC-end sequence forward and reverse pairs, but it was also the only one in which the majority of BES pairs spanned a distance that approximates the insert size of the papaya BAC library. In contrast, few papaya BES pairs mapped to within 300 kb in the *Arabidopsis* genome, and most of those spanned only 50 kb or less of *Arabidopsis* sequence. Seventeen paired papaya BAC ends could be mapped in both *Arabidopsis* and poplar, and thus enabled direct comparison of the corresponding regions from those two genomes relative to the papaya genome. Of these, 14 spanned an average 3.27 times larger region in poplar than *Arabidopsis* (extremes measured from 1.48–6.86 times larger). The small number of papaya BAC end sequences that co-mapped to the rice genome mostly spanned shorter than expected regions, although three of them spanned between 80 and 130 kb. Details of the number of BES pairs with BLAST matches in the target genome, mapped to the same chromosome, in the correct orientation and within 10–300 kb of each other are provided in Table 4.

To validate the methodology, several additional sets of BES pairs were mapped to the completed plant genomes. First, 102 *Arabidopsis* BAC end sequence pairs that met the same criteria were mapped onto the *Arabidopsis* genome. Ninety-four of these BAC end pairs were mapped successfully, and most of them (89) mapped to regions separated by between 80 and 140 kb, which approximates the average insert sizes (100 kb) of the TAMU (Choi et al. 1995) and IGF *Arabidopsis* BAC libraries (Mozo et al. 1998). Second, 8,525 *Brassica rapa* BES pairs were mapped to the *Arabidopsis* genome, and more than 50% of these pairs were localized within 300 kb of each other. Third, *Arabidopsis* and *Brassica* BES pairs were mapped to the more distantly related poplar genome. In this wider comparison, only about 1% of *Arabidopsis* and *B. rapa* BES pairs co-localized within 300 kb in the poplar genome. In contrast, 15.9% of papaya BES pairs (also order Brassicales) with homology to the poplar genome co-mapped within 300 kb, indicating a major difference in genome organization between members of the Brassicales and suggesting a

Table 4 Paired BAC ends statistics for different stages of comparative genome mapping

	No. of non-repeat BES pairs…	with BLAST matches in heterologous genome (1) and	on same chromosome or contig (2) and	in the correct orientation (3) and	within 300 kb (4)	Overall %	Average distance (kb) between paired BESs
Eurosids II versus eurosids I							
Papaya versus poplar	9,038	1,048 (11.6%)	283 (27.0%)	201 (71.0%)	167 (83.1%)	15.9	74.0
Brassica versus poplar	8,525	2,056 (24.1%)	253 (12.3%)	155 (61.3%)	20 (12.9%)	1.0	165.7
Arabidopsis versus poplar	3,648	633 (17.4%)	97 (15.3%)	56 (57.7%)	5 (8.9%)	0.8	230.0
Within Brassicales							
Papaya versus *Arabidopsis*	9,038	774 (8.6%)	276 (35.7%)	150 (54.3%)	53 (35.3%)	6.8	46.4
Poplar versus *Arabidopsis*	5,685	263 (4.6%)	84 (31.9%)	54 (64.3%)	19 (35.2%)	7.2	34.7
Within Brassicaceae							
Arabidopsis versus *Arabidopsis*	102	99 (97.1%)	98 (99.0%)	96 (98.0%)	94 (97.9%)	94.9	97.0
Brassica versus *Arabidopsis*	8,525	3,889 (45.6%)	2,790 (71.7%)	2,395 (85.8%)	2,049 (85.6%)	52.7	135.1
Euasterid versus eurosids							
Tomato versus *Arabidopsis*	10,394	408 (3.9%)	141 (34.6%)	83 (58.9%)	31 (37.3%)	7.6	45.3
Tomato versus poplar	10,394	470 (4.5%)	84 (17.9%)	54 (64.3%)	37 (68.5%)	7.9	117.3
Eudicot versus monocot							
Papaya versus rice	9,038	620 (6.9%)	109 (17.6%)	44 (40.4%)	11 (25.0%)	1.8	47.0

Low-copy BES pairs of five plant species were analyzed in sequential steps consisting of (1) mapping to optimal position in heterologous genome using BLAST, (2) identifying paired BAC ends that map to the same chromosome, (3) confirming proper orientation with respect to each other and (4) confirming localization to within 300 kb of each other. The percentage for each step is derived relative to the number of BES pairs from the previous stage. *Arabidopsis* BESs were mapped to the *Arabidopsis* genome as a control

higher degree of colinearity between papaya and poplar than within the Brassicales. This is confirmed by similar success rate of co-mapping papaya and poplar BES pairs in the *Arabidopsis* genome. Finally, tomato (order Solanales) BES pairs were mapped to both *Arabidopsis* and poplar with a similarly low success rate.

Significantly, the elevated level of co-mapping BAC end pairs observed in the *Arabidopsis–Arabidopsis*, *Brassica–Arabidopsis* and papaya–poplar comparisons correlates with the distance that separates the paired BAC ends in the heterologous genome (Fig. 2), indicating that they may represent truly colinear regions. Further support for colinearity comes from the fact that the proportion of BES pairs in the correct orientation is correlated with the phylogenetic relationship of the organisms, e.g. 85.8% of *Brassica* BES pairs are oriented correctly in the *Arabidopsis* genome as compared with 35% of papaya BES pairs. However, even the papaya–rice comparison yields higher than random (25%) BES pairs in the proper orientation. With only two exceptions (*Arabidopsis* and *Brassica*), paired BESs were separated by much less than an average BAC insert size (100 kb) in the *Arabidopsis* genome.

Discussion

Papaya BAC-end sequencing

More than 70% of BAC end sequence attempts yielded usable sequence data. A small percentage (3.9%) of BESs consisted entirely of vector sequence and likely originated from empty BAC clones. The estimate of empty clones for the two halves of the library obtained by restriction digest analysis of a relatively small number of

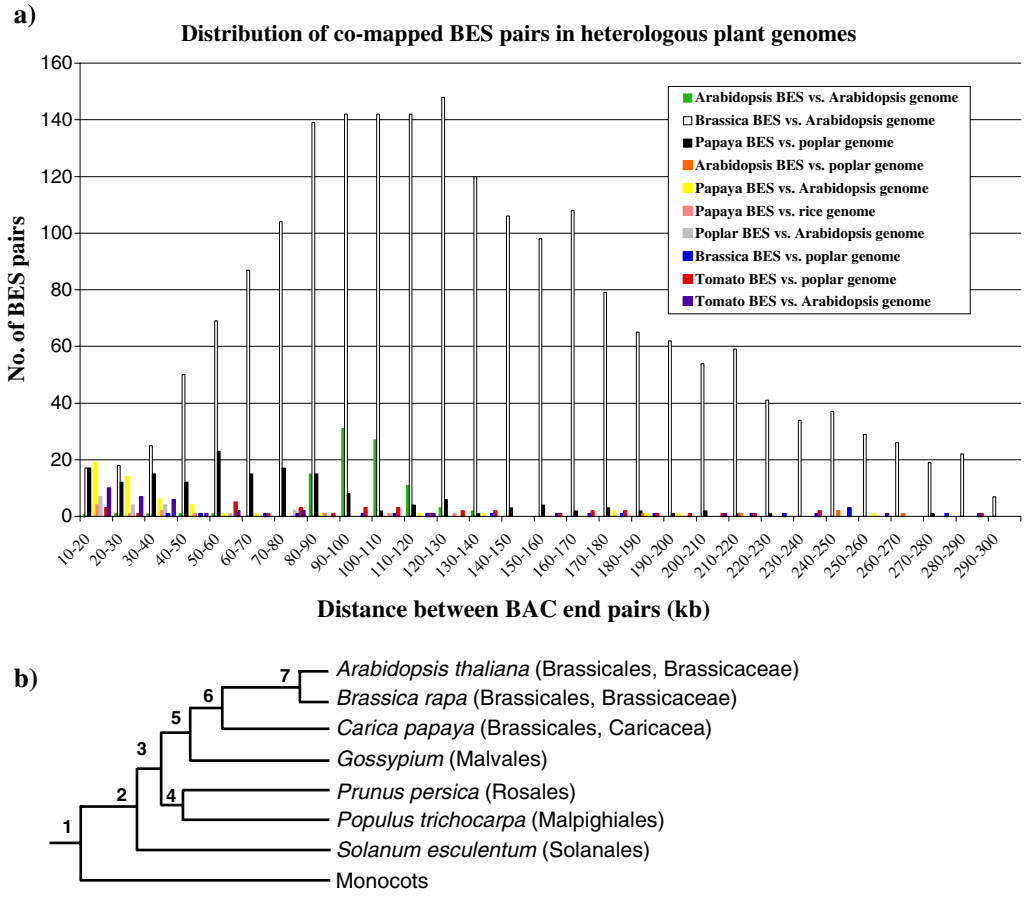

Fig. 2 Phylogenetic relationships do not always predict genome colinearity. **a** Distribution of BES pairs that are correctly oriented and located within 10–300 kb of each other in the heterologous genome. The number of BES pairs mapped to heterologous plant genomes is plotted against the distance separating the two BAC ends in the target genome. Low-copy BAC-end forward and reverse sequence pairs (9,038) of papaya were mapped to the *Arabidopsis thaliana, Populus trichocarpa*, and *Oryza sativa* genomes. Low-copy *Brassica rapa*, tomato and *Arabidopsis* BES pairs were mapped onto the *Arabidopsis* and poplar genomes. Poplar BESs generated in silico were processed identically and mapped onto the *Arabidopsis thaliana* genome to determine if co-mapped BES pairs represent colinear genomic regions. A set of low-copy *Arabidopsis* BESs was identically processed and compared with the *Arabidopsis thaliana* genome sequences as a control for the mapping process. Thirty-nine *B. rapa* BES pairs that mapped to poplar scaffold 6399 (chloroplast) were excluded. **b** Phylogenetic relationships of the species examined in **a**. Estimated times of divergence for nodes 1–7 are 161, 125, 109, 96, 90, 68–72 (Wikström et al. 2001), and 14–24 MYA (Yang et al. 1999), respectively

clones was 3.5 and 4.6% (Ming et al. 2001), which is in agreement with the BAC end sequence data. Ming et al (2001) also assayed the number of BACs derived from plastid DNA by hybridizing filters with the sorghum *rop* and *trunk* genes, and counted 504 positive clones among 36,864 colonies screened, or 1.37%. A BLAST homology search of the BESs under fairly stringent conditions revealed a comparable percentage (1.1%) of likely chloroplast-derived BACs.

Simple sequence repeats

Simple sequence repeats (or microsatellites) are a class of molecular marker that, due to their high level of polymorphism and simple assay, are widely used for generating genetic maps. The majority of papaya genetic maps constructed to date have employed less informative AFLP markers (Kim et al. 2002; Van Droogenbroeck et al. 2002; Ma et al. 2004). The 7,456 potential SSR markers discovered in this project constitute a welcome addition to the papaya breeder's toolbox. In order to promote use of these newly discovered markers, we provide computationally derived flanking primers that can be used to PCR-amplify 2,575 of the SSRs discovered in this project at (http://www.genomics.hawaii.edu/cgi-bin/papaya/BES/ssrNP.cgi).

The most abundant SSRs in all four size categories were AT-rich, i.e. poly(A) and poly(T) (representing 94.7% of all homopolymers of length ≥ 12); poly(AT) and poly(TA); poly(AAT) and poly(ATT); and poly(AATT)/poly(TTAA). This is in agreement with previous reports of microsatellite abundance in other species: poly(AT)/(TA) and AT-rich trinucleotide repeats were the most abundant repeats of their class in *A. thaliana* and yeast (Katti et al. 2001). However, as Temnykh et al. (2001) point out, the majority of poly(AT) repeats in rice do not amplify cleanly, possibly because most lie in non-coding regions and are frequently associated with repeat elements. In agreement with the findings in rice (Temnykh et al. 2001), where AT-rich di- and trinucleotide repeats are overrepresented in intergenic regions, these repeats are underrepresented in papaya BESs that have cDNA homologies.

Longer microsatellites, also known as class I (at least 20 nt) are more likely to be hypervariable and thus polymorphic. Dinucleotide repeats in general, and specifically poly(AG) and poly(GA) repeats, were among the longest found in papaya, the longest being 70 bases of pure poly(AG). Most tetranucleotide repeats (95.6%) are less than 20 bases long and thus unlikely to be highly polymorphic. Consistent with the data of Katti et al. (2001), the frequencies of all mono-, di-, tri-, and tetranucleotide tandem repeats were found to decrease exponentially with increasing microsatellite length (Fig. 1), which may be an indication of selective pressures against very long tandem repeats.

Similar to what was observed for Rosaceae ESTs (Jung et al. 2005), dinucleotide repeats represent the most abundant of the four microsatellite classes. Trinucleotide tandem repeats are least abundant in papaya BAC-end sequences, but those that do not consist entirely of AT nucleotides are enriched in BESs with cDNA homology. This enrichment of trinucleotide repeats in putative coding regions of ESTs was also observed for the Rosaceae (Jung et al. 2005), and may be due to a higher tolerance to trinucleotide repeats in protein coding regions, since they do not shift the open reading frame during expansion or contraction (Katti et al. 2001). However, disruption of protein structure may lead to the suppression of the trinucleotide tandem repeats (Temnykh et al. 2001). With the exception of trinucleotide repeats, all SSRs were overrepresented in the BESs that have no homology to the cDNA or repeat databases.

Plant repeat elements

Known plant repeats were discovered in more than 16% of the papaya BESs, and more than 90% of these repeats are transposons. The transposon content of papaya thus appears to be intermediate between that of rice (35%, International Rice Genome Sequencing Project 2005) and *Arabidopsis* (10%, The Arabidopsis Genome Initiative 2000). However, the detection of repeat elements discovered in papaya BESs may be biased significantly by the restriction enzyme used to generate the BAC clones. Several novel repeats were identified in the papaya BESs, which appear to account for a significant portion of the papaya genome. The intermediate size of the papaya genome compared with that of rice (430 Mbp) and *Arabidopsis* (125 Mbp) may, in part, be accounted for by these repeat elements.

Most known plant repeat elements that have been found in other organisms were also found in papaya BESs. Retrotransposons, which were found in 13.5% of all papaya BESs, are the most abundant repeat elements. The ratio of *Ty3-gypsy* to *Ty1-copia* retrotransposons in papaya BESs is 1:1 and resembles that of the *Arabidopsis* genome (The Arabidopsis Genome Initiative 2000) and maize BESs (Messing et al. 2004). In contrast, this ratio in the rice genome was reported to be around 2:1 (International Rice Genome Sequencing Project 2005).

Gene content of papaya BESs

A total of 6,769 BESs (or 19.1% of the total) exhibited sequence homology to the *Arabidopsis* cDNAs and could be used to rapidly clone candidate genes from the BAC library. Based on an estimated genome size of 372 Mbp for papaya, and an average gene length (2 kb) similar to that of *Arabidopsis* (The Arabidopsis Genome Initiative 2000), the 19.1% BESs with cDNA homology suggest a total cDNA coding capacity of 71.1 Mbp and a total gene content of 35,526. Combining the SSR with cDNA data allows selection of SSRs from gene-rich regions, which may be particularly useful for mapping studies. Most protein homologies are accounted for by the *Arabidopsis* cDNA database. Only 1.5% of the high-quality BESs had a match in the NCBI *nr* database but not in the cDNA, RefSeq or plant repeat databases.

Papaya-specific repeats

Clustering of the BAC end sequences without matches in any of the four databases examined, revealed several known (mostly sex chromosome-specific) and novel papaya-specific repeats. Examination of the 10 largest clusters, which account for 1,513 BESs (or 4.3% of the total number of high-quality BESs) revealed that six of these papaya-specific repeats have homology to known sex-linked markers from the papaya Y chromosome; these BACs may have originated from that chromosome. In addition, four previously unknown papaya-specific repeats were identified and their addition to the *Plant Repeat Database* should be considered. Fluorescent in situ hybridization experiments may reveal interesting genomic distribution patterns for these novel repeats.

Comparative genome mapping

Forward and reverse BES read pairs represent sequence tags separated by the length of a BAC (generally around 100 kb or more). When these tags consist of single- or low-copy sequences, they can be mapped unambiguously to completely sequenced heterologous genomes to establish a measure of microsynteny or colinearity. Microsynteny has been well documented between member species of the same plant families (e.g. Solanaceae, Poaceae), where it has been exploited to identify candidate genes in the heterologous species (Yan et al. 2003; Huang et al. 2005). Despite the fact that detection of colinearity is complicated by genome duplications and subsequent gene loss in the lineages leading to both of these organisms, limited colinearity has been observed even between plants from different orders, such as cotton and *Arabidopsis* (Rong et al. 2005), which are estimated to have diverged approximately 90 MYA (Fig. 2b; Wikström et al. 2001; Rong et al. 2005). Zhu et al. (2003) detected highly degenerate microsynteny between *Medicago* and *Arabidopsis*, which are as distantly related as poplar and *Arabidopsis* (last common ancestor ca 109 MYA—see Fig. 2b).

We mapped non-repeat-containing papaya BES pairs to the three completely sequenced plant genomes (*A. thaliana*, *Populus trichocarpa* and *Oryza sativa*) to determine the extent of colinearity between papaya and these species. BES pairs that are mapped to within 300 kb of each other and oriented correctly in the heterologous genome were considered to span potentially colinear regions. We validated our computational pipeline and identified its limitations by attempting to map 102 *Arabidopsis* BES pairs to the *Arabidopsis* genome. Poor sequence quality of four BESs prevented the mapping of three of the pairs. Paired reads were mapped in the correct orientation and within a 10–300 kb range 95% of the time. In two cases BESs of the same pair were located < 10 kb apart, in another case one BES was mismapped due to multiple matches of equal or better E value to a second locus and in two cases the paired reads were in the wrong orientation. Most of these errors appear to be due to tracking errors associated with these earliest of plant BES data sets.

We next mapped *B. rapa* BESs against the *Arabidopsis* genome. Both of these species are in the Brassicaceae family (order Brassicales) that diverged an estimated 14.5–20.4 MYA (Yang et al. 1999). Almost 53% of all *B. rapa* BES pairs that could be mapped to the *Arabidopsis* genome were located in the correct orientation and within 300 kb of each other.

Papaya is a member of the sister family Caricaceae, which is a basal family within the order Brassicales. As expected, papaya BESs mapped to the *Arabidopsis* genome at a lower frequency (6.8%) than the *B. rapa* BES pairs. This drop in colinearity suggests at first glance that significant colinearity between species is limited to members of the same family. However, a comparison of papaya BES pairs with the more distantly related poplar genome (order Malpighiales) yielded a higher level of colinearity (15.9%). This higher proportion of BES pairs mapped to poplar as compared with *Arabidopsis* is not due to a novel repeat element shared by papaya and poplar, as more than 90% of the 334 BESs mapped to the poplar genome have homologs in the *Arabidopsis* cDNA (290 BESs) or *nr* (8 BESs) database. Furthermore, the region spanned by the paired BAC ends in poplar (10–160 kb) was more similar to the papaya BAC insert size than the generally shorter (10–50 kb) *Arabidopsis* region (Fig. 2). This fact, which may be due to genome reduction in *Arabidopsis*, plus the fact that the number of co-mapped pairs for *Arabidopsis* are significantly lower, suggest that, of the three completed plant genomes currently available, the poplar genome may be the most suitable for comparative genomics with papaya.

The fact that more papaya BACs appear to be colinear with poplar than with *Arabidopsis* is somewhat surprising, given current understanding of the taxonomy of these species: both papaya and *Arabidopsis* are members of the order Brassicales in the eurosids II group, whereas poplar is a member of the Malpighiales in the eurosids I group (Judd et al. 2002). Thus, a higher number of colinear BES pairs would be expected for the papaya–*Arabidopsis* than papaya–poplar comparison. However, the basal position of the Caricaceae family within the order Brassicales (Judd et al. 2002; Fig. 2b), may partially account for the apparently higher colinearity of papaya with poplar.

The most important factor in the reduced colinearity between papaya and *Arabidopsis* may be a recent (predating the divergence from *Brassica*) duplication of the *Arabidopsis* genome that was followed by loss of approximately 70% of the duplicated genes (Bowers et al. 2003). As evidenced by the small genome size of *Arabidopsis*, this lineage may be unusually adept at eliminating duplicated genes and may thus represent a particularly unsuitable genome to use as a template for comparative genomics. Assuming random loss, the probability of preserving both members of a progenitor BES pair in the *Arabidopsis* genome would be 0.3^2 or 9%, close to the 7% observed. Note that this assumes no similar event (i.e. genome duplication followed by gene loss) in papaya. Large-scale genome duplication in the past 70 million

years has been documented for many crop species, including grasses, tomato, potato and soybean (Schlueter et al. 2004). In this context it is noteworthy that sex chromosomes, such as the one that appears to be evolving in papaya and other Caricaceae (Liu et al. 2004), present a roadblock to polyploidization and may thus have acted to maintain a more ancestral genome in papaya.

While *Arabidopsis*, *Brassica* and *Carica* are all rosids, *Lycopersicon* is a member of the asterids (order Solanales). The last common ancestor of tomato and the Brassicales existed approximately 125 MYA. Tomato BES pairs, when mapped to the *Arabidopsis* and poplar genomes, co-localized to within 300 kb at similar rates (7.6 and 7.9%, respectively).

The overall amount of colinearity appears to be quite low in all comparisons: even in *Populus trichocarpa*, which showed the highest number of co-mapped papaya BAC end pairs, only 1.8% of all "low-copy" BAC end pairs map to within 300 kb of each other. This is largely due to the fact that between 88.4% (poplar) and 93.1% (rice) of papaya BES pairs could not be mapped unambiguously using BLAST, presumably in part because the BES consisted of a species-specific repeat that was not present in the plant repeat databases used (note, for example, the large number of novel repeats discovered in papaya).

In all heterologous genomes the proportion of paired BAC ends that mapped to the same chromosome and within 300 kb of each other exceeded that expected by random chance. For example, in *Arabidopsis* (with five chromosomes), the probability of any two random BAC ends mapping to the same chromosome is 20.55% (adjusted for the variation in chromosome size), yet for more than 35% of BAC end pairs both ends mapped to the same chromosome. The orientation of the paired BAC ends mapping to the same chromosome, if random, is expected to be correct 25% of the time. This value is exceeded in all comparisons, but more closely related species exhibit higher values (e.g. *Brassica–Arabidopsis* = 86%). When mapped to the poplar genome, the papaya BES pairs are oriented correctly in 10% more cases than paired BESs of the Brassicaceae. Furthermore, in poplar more than 83% of paired papaya BAC ends that map to the same chromosome and are oriented correctly lie within 300 kb of each other, even though the probability of any two BESs lying within this distance is only 1.85% (adjusted for the differing sizes of the 19 linkage groups). In contrast, Brassicaceae BES pairs map to the same 300 kb poplar interval in only about 10% of cases—still above random levels but far less than the 83% observed for papaya.

We emphasize that the co-mapping of paired BESs to within 300 kb in the heterologous genome does not necessarily reflect complete colinearity of all intervening sequences. However, the high proportion of co-mapped BAC ends in the correct orientation, as well as the correlation of the distance between co-mapped BAC ends with the size of the heterologous genome (compare tomato vs. *Arabidopsis* and tomato vs. poplar) support the notion that these regions may indeed be largely colinear. The availability of only three completely sequenced and distantly related plant genomes limits the number of comparisons in which this can be examined in detail. Closer examination of the 19 potentially colinear poplar–*Arabidopsis* regions revealed only limited and degenerate colinearity, similar to what was observed for peach–*Arabidopsis* comparisons by Georgi et al. (2003).

Remarkably, three peach BACs (GenBank accession numbers AC154900.1, AC154901.1 and AF467900.1) showed extensive colinearity with the poplar genome (supplementary Fig. 1). Peach (*Prunus persica*) and poplar shared a common ancestor at 94–98 MYA (Wikström et al. 2001 and Fig. 2b) and, if these three BACs are representative of the peach genome, colinearity between these two species is high and poplar will serve as an excellent template on which to assemble peach genome sequence. Definitive evaluation of the exact extent of colinearity between papaya and poplar will have to await the generation of long contiguous stretches of papaya genome sequence.

We show that significant microsynteny can be detected beyond the family level and is not always congruent with our current understanding of phylogenetic relationships. In fact papaya, a basal member of the Brassicales, exhibits a higher level of apparent colinearity with the poplar (est. divergence 109 MYA) than with the *Arabidopsis* genome (est. divergence 70 MYA), suggesting that the *Arabidopsis* genome has been subjected to more extensive rearrangements, most likely to the documented recent α duplication event followed by large-scale gene loss. However, factors, such as long generation time and mating strategy (self-pollinators vs. outcrossers), may contribute to the higher apparent colinearity we observed between the papaya and poplar genomes. Inbreeding species, such as *Arabidopsis*, may be significantly more tolerant to major genome rearrangements (e.g. Robertsonian translocations) than obligate outcrossers.

The depth of sequence coverage that must be obtained for each plant family thus depends in part on the reproductive biology of its members. In general, it might be prudent to sequence to completion one member of each plant family, and generate BAC frameworks for other family members that can be used to clone species-specific candidate genes. This may be particularly effective in the Rosaceae family, which includes many fruit tree crops. In some cases, where genome rearrangements occur at significantly slower rates than in *Arabidopsis*, a complete genome from a sister family may suffice for comparative genomics purposes.

Conclusion

The papaya BAC end sequences described here reveal a plant genome containing all of the major repeat classes, a gene set similar to that of *A. thaliana*, large numbers of useful microsatellites, several novel papaya-specific

repeats and a surprising amount of colinearity with the poplar genome. These sequences will provide a valuable resource for future physical mapping or whole genome sequencing efforts. The microsatellite markers discovered in these BESs will accelerate the construction of a new generation of genetic maps.

Acknowledgements The Center for Genomics, Proteomics and Bioinformatics Research Initiative at the University of Hawai'i contributed 11,013 of the BAC end sequence chromatograms that are described here, using funds provided by the University of Hawai'i. The remainder of the BES data was generated with funds from the United States Department of Agriculture (USDA) Tropical and Subtropical Agriculture Research Program (grant HAW00557G) and a USDA-Agricultural Research Service Cooperative Agreement (CA 58-3020-8-134) with the Hawai'i Agriculture Research Center.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9(3):211–215

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu J, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo M, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:1–10

Cheng Z, Presting G, Buell CR, Wing RA, Jiang J (2001) High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics 157:1749–1757

Choi S, Creelman RA, Mullet JE, Wing RA (1995) Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*. Plant Mol Biol Rep 13:124–129

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wend MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Georgi LL, Wang Y, Reighard GL, Mao L, Wing RA, Abbott AG (2003) Comparison of peach and *Arabidopsis* genomic sequences: fragmentary conservation of gene neighborhoods. Genome 46:268–276

Goff SA, Ricke D, Lan T, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange B, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun W, Chen L, Cooper B, Park S, Wood T, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller R, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296:92–100

Hong CP, Lee SJ, Park JY, Plaha P, Park YS, Lee YK, Choi JE, Kim KY, Lee JH, Lee J, Jin H, Choi SR, Lim YP (2004) Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. Mol Genet Genomics 271:709–716

Huang S, van der Vossen EAG, Kuang H, Vleeshouwers VGAA, Zhang N, Borm TJA, van Eck HJ, Baker B, Jacobsen E, Visser RGF (2005) Comparative genomics enabled the isolation of the *R3a* late blight resistance gene in potato. Plant J 42:251–261

Ilic K, SanMiguel PJ, Bennetzen JL (2003) A complex history of rearrangements in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci USA 100:12265–12270

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (2002) Plant systematics: a phylogenetic approach, 2nd edn. Sinauer Associates, Inc. Sunderland

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics 5:136–143

Katti M, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

Kim MS, Moore PH, Zee F, Fitch MM, Steiger DL, Manshardt RM, Paull RE, Drew RA, Sekioka T, Ming R (2002) Genetic diversity of *Carica papaya* as revealed by AFLP markers. Genome 45:503–512

Lange BM, Presting G (2004) Genomic survey of metabolic pathways in rice. In: Romeo JT (ed) Recent advances in phytochemistry. Elsevier, Amsterdam, pp 111–137

Liu Z, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu Q, Pearl HM, Kim MS, Chartton JW, Stiles JI, Zee FT, Paterson AH, Ming R (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature 427:348–352

Ma H, Moore PH, Liu Z, Kim MS, Yu Q, Fitch MM, Sekioka T, Paterson AH, Ming R (2004) High-density linkage mapping revealed suppression of recombination at the sex determination locus in papaya. Genetics 166:419–436

Mao L, Wood T, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res 10:982–990

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci 101:14349–14354

Ming R, Moore PH, Zee F, Abbey CA, Ma H, Paterson AH (2001) Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. Theor Appl Genet 102:892–899

Mozo T, Fischer S, Shizuya H, Altmann T (1998) Construction and characterization of the IGF *Arabidopsis* BAC library. Mol Gen Genet 258:562–570

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of Brassica olearacea var. alboglabra that are homeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J 23:233–43

Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity and evolution of rice chromosome 10. Science 300:1566–1569

Rong J, Bowers JE, Schulze SR, Waghmare VN, Rogers CJ, Pierce GJ, Zhang H, Estill JC, Paterson AH (2005) Comparative genomics of *Gossypium* and *Arabidopsis*: unraveling the consequences of both ancient and recent polyploidy. Genome Res 15:1198–1210

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols (methods in molecular biology). Humana Press, Totowa, pp 365–386

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47:868–877

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA 89:8794–8797

Temnykh S, DeClerck G, Lukashova A, Lipoviich L, Cartinhour, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length

variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

The Arabidopsis Genome Initiative (2000) Analysis of the genome structure of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Tomkins J, Fregene M, Main D, Kim H, Wing R, Tohme J (2004) Bacterial artificial chromosome (BAC) library resource for positional cloning of pest and disease resistance genes in cassava (*Manihot esculenta* Crantz). Plant Mol Biol 56:555–561

Van Droogenbroeck B, Breyne P, Goetghebeur P, Romeijn-Peeters E, Kyndt T, Gheysen G (2002) AFLP analysis of genetic relationships among papaya and its wild relatives (*Caricaceae*) from Ecuador. Theor Appl Genet 105:289–297

Wikström N, Savolainen V, Chase M (2001) Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B 268:2211–2220

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA 100:6263–6268

Yang Y-W, Lai K-N, Tai P-Y, Li W-H (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. J Mol Evol 48:597–604

Zhao S, Shatsman S, Ayodeji B, Geer K, Tsegaye G, Krol M, Gebregeorgis E, Shvartsbeyn A, Russell D, Overton L, Jiang L, Dimitrov G, Tran K, Shetty J, Malek JA, Feldblyum T, Nierman WC, Fraser CM (2001) Mouse BAC ends quality assessment and sequence analyses. Genome Res 11:1736–1745

Zhu H, Kim D-J, Baek J-M, Choi H-K, Ellis LC, Küester H, McCombie WR, Peng H-M, Cook DR (2003) Syntenic relationships between *Medicago trunculata* and *Arabidopsis* reveal extensive divergence of genome organization. Plant Physiol 131:1028–1026